Network analysis methods for smart inspection in the transport domain

Bruin, G.J. de

4

Understanding dynamics of truck co-driving networks

In this chapter, we move from the investigation of a generic network science problem towards the transportation domain by investigating the *behavior* of trucks and their drivers using a link prediction approach. Social links may exist between trucks, e.g., because their drivers work for the same company. We call the process where two trucks follow the same route at the same time *co-driving* (Definition 12). It means that the trucks are potentially *socially linked*.

Understanding *truck co-driving behavior* is important because co-driving can have a positive environmental impact. We aim to increase our understanding and will investigate the so-called *co-driving network*, extracted from a spatiotemporal dataset encompassing millions of truck measurements passing eighteen different highway locations in the Netherlands. It leads us to Research question 3, which reads as follows.

Research question 3: *How do network structure and vehicle attributes relate to co-driving behavior?*

We explore a *link prediction* approach to understand the (social) processes underlying the co-driving behavior. By investigating the importance of different types of features (e.g., vehicle attributes) provided to the link prediction algorithm, we learn step by step the relation between network structure and co-driving behavior.

The current chapter corresponds to the following publication:
G. J. de Bruin, C. J. Veenman, H. J. van den Herik, and F. W. Takes. „Understanding dynamics of truck co-driving networks." In: *Proceedings of the 8th International Conference on Complex Networks and Their Applications*. Studies in Computational Intelligence 882. Springer, 2020, pages 140–151. DOI: 10.1007/978-3-030-36683-4_12

4.1 Co-driving network

In the literature, many published studies concerning social network analysis use spatiotemporal data. This often allows enriching the analysis with meaningful insights into social processes. Much of the research performed so far used either GPS [44, 142], WiFi [171] or calls from mobile phones [196] to study social processes. In this study, we will analyze 19 million truck movements.

The goal is to study social phenomena among truck drivers to understand why truck drivers are engaged in so-called *co-driving behavior* with other drivers. In simple terms, co-driving is the activity where two trucks drive together, i.e., are frequently at the same place simultaneously. Here we assume a direct and natural relation between a truck and its driver, meaning that a truck driver only drives one truck and the same driver always drives the truck. Some strict selection criteria ensure that only *intentional* (or similarly, *systematic*) co-driving activity is investigated (see also Definition 12). The criteria are explained in Subsection 4.4.3.

Co-driving behavior is known to have a potentially positive impact on the environment through optimizing logistics and consequently reducing fuel use [188]. Hence, an improved understanding of co-driving behavior may stimulate co-driving behavior. Moreover, innovative forms of transportation, such as autonomous driving, may have significant implications for this behavior.

We construct a so-called *co-driving network* from the data at our disposal. The nodes of the network are trucks. A link is made when the two trucks frequently show intentional co-driving behavior. Other related work on similar data will focus on communities and static properties of the co-driving network, see Chapter 5 and [30].

This chapter aims to learn the relation between *the structure of the co-driving network* and *vehicle characteristics*. To this end, we use a link prediction approach [96]. More precisely, we develop a machine learning classifier that predicts whether two nodes that are so far unconnected, do connect. We then use a future snapshot of the network to check whether the pair of nodes did connect. Subsequently, we investigate the feature importance of each type that occurs in the link prediction classifier. The measure of importance allows us to understand what is assessed as important by the classifier, and thus what aspects are contributing to co-driving behavior. The features used can be categorized into four different types of features.

1. *Neighborhood* features relate to the local embedding in the co-driving network.
2. *Node* features relate to static meta-information of trucks.
3. *Path* features describe distance-related properties of truck pairs based on the global structure of the network.
4. *Spatiotemporal* features consider locations and periods.

The overall structure of this chapter coincides with the research methodology (see Section 1.7). We start with the introduction of the co-driving network in Section 4.1. In

Section 4.2, relevant work is provided on analyzing dynamics in social networks, including spatiotemporal data. Section 4.3 describes the spatiotemporal truck data. Section 4.4 reports how a co-driving network is constructed from the data. In this section, we also discuss the characteristics of the obtained network. Section 4.5 provides our research methodology for the experiments at hand, i.e., a formal description of the link prediction approach. It also explains how the different features are constructed from both the data and the obtained network. Section 4.6 outlines the experimental setup, demonstrates the performance of the link prediction approach, and assesses the importance of the features. Finally, in Section 4.7 we arrive at the conclusions and suggestions for future work.

4.2 Relevant related work on dynamics in networks

From the substantial body of related work on spatiotemporal data, we have selected three approaches frequently used to study dynamics in networks at the level of individual nodes. These three different approaches have in common that they all try to understand the underlying social network by studying node attributes available in the data.

First, Sekara *et al.* use sensors to measure proximity of students [175]. The authors show that when high-resolution data is available (both in time and location), groups of interacting nodes can be observed instantaneously. Hence, making sense of individual node attributes using network measures can be performed directly. For example, the authors show that the students may explore new locations in groups during the weekend, while the groups tend to be at the exact location.

Secondly, Kossinets and Watts analyze e-mail data gathered from students and employees at a university [100]. Unlike our truck data, e-mail data does not contain spatial information. In contrast and as an addition, this work collects and analyzes different node attributes such as professional status, gender, and age.

Finally, Wang *et al.* analyze the mobility patterns by tracking the mobility and interactions of millions of mobile phone users [196]. A social network is constructed from phone calls, where users are connected when they communicate. Three contributions from this literature are mentioned below.

1. The authors have established that spatial trajectories of two users strongly correlate when they are close in the social network.
2. Mobility features have a high predictive power concerning which nodes will connect; the prediction power is comparable to the power of network proximity features.
3. Link prediction performance can be improved by exploiting network proximity and mobility features.

Here, we remark that we have used a similar link prediction approach in our work. In addition, we have adapted findings from other related works [100, 175] by (1) distinguishing between weekends and weekdays and (2) using both network and static attributes.

4.3 Truck mobility data

Data collection of truck mobility data occurred at eighteen different locations throughout the Netherlands between 2016 and 2018. Every truck passing these locations is registered using an ANPR system. The data is obtained by the same systems as used in Chapter 5. At some locations, the registration systems faced an unexpected downtime. Only registrations from six out of eighteen systems have been considered to ensure a sufficiently valid range of data. These systems are located near the port of Rotterdam. Furthermore, registrations with low-quality data have been removed, such as (1) invalid characters in license plates and (2) non-existing countries.

We remark that the aforementioned quality selections have reduced the total number of registrations from 18,678,420 to 9,202,764. The monthly variation in truck registrations is provided in Figure 4.1, where we show for each of the 25 months (from January 2016 to February 2018) how many trucks are registered. We remark that the number of registrations after applying the quality selections is more stable over time. In Figure 4.2 the histogram of the number of registrations per truck is shown (note that both axes have logarithmic scales). For example, we see that about 1 million trucks are registered only once. More importantly, we see that the distribution of the number of registrations per truck remains similar after data selection.

4.4 The co-driving network

In Subsection 4.4.1, we start with three relevant concepts and two criteria to arrive at a procedure to obtain intentional truck co-driving events. Then we describe how the co-driving network is constructed from these events in Subsection 4.4.3. Subsection 4.4.4 continues with statistics of the acquired network to compare these to other social networks.

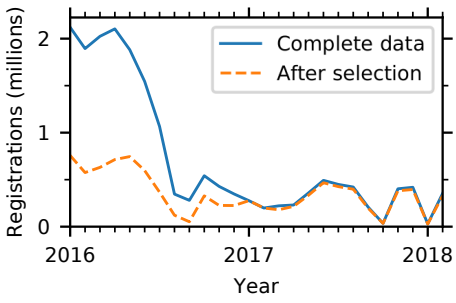


Figure 4.1: Monthly variations in truck registrations.

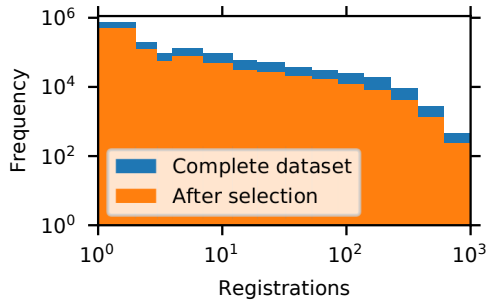


Figure 4.2: Histogram of number of registrations per truck.

4.4.1 Procedure to obtain intentional co-driving events

We will now provide the procedure to obtain intentional co-driving events (see Definition 12) with the help of three relevant concepts: (1) dataset of all registrations, (2) a co-driving event, and (3) an intentional co-driving event.

Concept 1. Our *dataset of all registrations* (as mentioned in Section 4.3) is denoted by $\mathcal{D}$. We use $\mathcal{D}_u$ to refer to all registrations x_i in dataset $\mathcal{D}$ from truck u with license plate $lp_i = u$. More formally, $\mathcal{D}_u = \{x_i \in \mathcal{D} : lp_i = u\}$.

Concept 2. A *co-driving event* (u, v, t_i) happens when two registrations $x_i \in \mathcal{D}_u$ and $x_j \in \mathcal{D}_v$ from trucks u and v exist with the same location $loc_i = loc_j$ at time t_i provided that they have at most $\Delta t = t_j - t_i$ (with $t_j < t_i$) seconds between them (see Subsection 4.5.2).

Concept 3. A co-driving event may occur *randomly* or *intentionally*. The following two criteria ensure that only intentional co-driving events are studied.

Criterion 1: Sufficient small time interval. The two registrations $x_i \in \mathcal{D}_u$ and $x_j \in \mathcal{D}_v$ from trucks u and v should exist with at most $\Delta t \leq \Delta t_{\text{max}}$ seconds apart. (Seconds will be further in this thesis be abbreviated by s.) In Subsection 4.4.2, we will briefly discuss why we set the Δt_{max} parameter to $8\,\mathrm{s}$. It ensures that trucks should be sufficiently close to each other when intentionally co-driving by setting a maximal time interval between two co-driving trucks.

Criterion 2: At least two separate co-drive events. To prevent a random co-driving event is marked as an intentional co-driving event, we require at least two separate co-driving events between trucks u and v. Moreover, these two separate events should occur with at least two hours difference, i.e., two co-driving events exist, (u, v, t_i) and (u, v, t_j), for which holds that $|t_i - t_j| \geq 2\,\mathrm{h}$. With the latter requirement, we ensure that the two co-driving events originate from different truck journeys (we assume that in $2\,\mathrm{h}$, trucks are either outside the Netherlands or driving on the next journey).

4.4.2 Determining maximal time interval between co-driving trucks

In Criterion 1 above, we introduced the Δt_{max} parameter, determining the maximal time interval between two co-driving trucks. We mentioned that $\Delta t_{\text{max}} = 8\,\mathrm{s}$ is deemed appropriate. We will now explain why.

There is a trade-off. High values will select a large share of random co-driving events, while low values will omit intentional co-driving behavior. We present three considerations when determining the value of Δt_{max}.

Consideration 1. Figure 4.3 shows the distribution of the time gap between two co-driving events. On the horizontal axis, we see the time gaps in whole seconds; the vertical axis denotes the relative frequency of that time gap (altogether the frequencies add up to 1). We note that distinct behavior is shown for random (yellow) and

intentional (blue) co-driving events. Intentionally co-driving trucks drive closer together than randomly co-driving trucks. We further note that the time gap in intentional co-driving trucks peaks at around $\Delta t = 2\,\mathrm{s}$ and is close to the $\Delta t = 1.3\,\mathrm{s}$, which is considered a minimum safe driving gap between two trucks [116]. After $\Delta t = 8\,\mathrm{s}$ the relative frequency of intentional co-driving trucks becomes similar to that of randomly co-driving trucks. This may indicate that only random co-driving events are selected as intentional co-driving from this value onward.

Consideration 2. Figure 4.4 shows the distribution of the number of trucks driving between two trucks involved in intentional co-driving for various values ($\Delta t_{\max} = 4,\ 8,\ 16$ and $32\,\mathrm{s}$). The horizontal axis denotes the number of trucks, and the vertical axis the cumulative relative frequency of that number of trucks driving between the co-driving pair. For values between $\Delta t = 4$ and $8\,\mathrm{s}$, we observe that virtually all trucks drive with at most one truck between them. Higher values result in a non-negligible probability that more than two trucks are driving between the two co-driving trucks. It is unlikely that trucks are intentionally co-driving when more than two trucks drive between these trucks because it is harder to coordinate routing. This is the case for values of $\Delta t_{\max} = 16\,\mathrm{s}$.

Consideration 3. We rationalize that *intentionally following* a truck is only possible when a maximum of a couple hundred meters between the two trucks exists. Provided that trucks in our data drive typically at a speed of around $20\,\mathrm{m\,s^{-1}}$, reasonable values for $\Delta t_{\max}$ should be at most $20\,\mathrm{s}$ to $30\,\mathrm{s}$.

The considerations above have led us to properly select intentional co-driving behavior for further analysis in this chapter.

4.4.3 Network construction

After applying the two criteria to select intentional co-driving events, the temporal network $G = (V, E)$ is constructed. In this network, the nodes are the trucks $u, v \in V$ that frequently show intentional co-driving behavior (have at least one edge). The links of

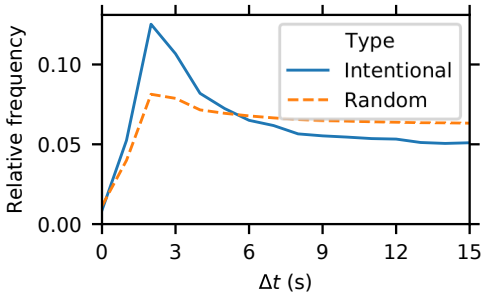


Figure 4.3: Frequency distribution of Δt for both intentional and random co-driving.

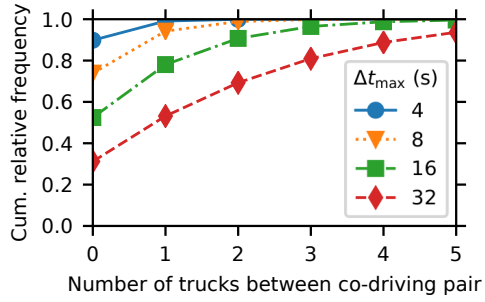


Figure 4.4: Number of trucks driving between a pair of co-driving trucks.

this network consist of the obtained co-driving events $(u, v, t_i) \in E$ between those trucks. We note that multiple links (u, v, t_i) exist between two nodes u and v with different t_i due to Criterion 2 (see Subsection 4.4.1) to select only intentional co-driving. We refer to the number of links between u and v as $w_{u,v}$, with $w_{u,v} \geq 2$ as a result of the two criteria discussed above. When no links exist between u and v, the weight $w_{u,v}$ equals 0.

4.4.4 Network statistics

In Table 4.1, we summarize nine statistical properties calculated from our obtained network. All these statistics are explained in Section 1.2. The degree distribution of *each* node is shown in Figure 4.5a. We show the node strength distribution in Figure 4.5b. The vertical axis denotes the frequency of the (a) number of neighbors (degree) and (b) *node strength* of all nodes in the truck co-driving network. The node strength of a node is equal to the sum of the weights of the nodes connected to that node.

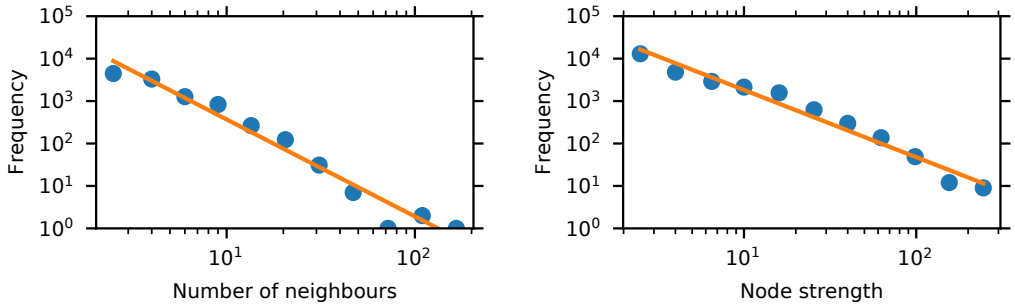
Our network is remarkably similar to other (social) networks. We find the following common properties [10, 12, 197] (see Section 1.2).

- A *Giant Component* is present that spans most nodes and edges (cf. item 4 in Subsection 1.2.4).
- *Sparseness* of edges, with only 0.2‰ of possible pairs of nodes being connected (cf. item 5 in Subsection 1.2.4).
- *Power-law behavior* in both the degree and weight distribution as seen in Figures 4.5a and 4.5b (cf. item 7 in Subsection 1.2.4).
- A relatively *low average path length* (cf. item 6 in Subsection 1.2.4).

Because our network is remarkably similar to other networks, we may conclude that the network construction is successful. In Sections 4.5 and 4.6, we will search for complex relationships between truck drivers that can be understood by investigating the obtained truck co-driving network.

Table 4.1: Nine statistical properties of the co-driving cargo truck network.

Property	Value
Number of nodes	$25,553$
Number of links	$73,059$
Number of connected node pairs	$27,986$
Fraction nodes in Giant Component	62%
Fraction links in Giant Component	79%
Density	2.2×10^{-4}
Power law exponent γ	3.3
Average shortest path length	7.8
Diameter	24



(a) Degree distribution.

(b) Strength distribution.

Figure 4.5: (a) Degree and (b) strength distribution of co-driving cargo truck network. (Note the logarithmic axes.)

4.5 Chapter research methodology

This section presents the methodology used in this chapter for the analysis of the dynamics of the co-driving network. We start with a description of the proposed link prediction approach in Subsection 4.5.1. The features are provided in Subsection 4.5.3. In Subsection 4.5.4 we discuss the setup of the classifier. Finally, we provide the measures taken to reduce the observed class imbalance in Subsection 4.5.4.

4.5.1 Link prediction

We start by describing link prediction (see also Definition 10). We tailor similar notations used in the Chapters 2 and 3 to the problem at hand.

The link prediction problem is as follows. Given a network observed at a time interval $[t_a, t_b]$ (with $t_a < t_b$), the link prediction classifier needs to predict newly formed links in the network at an evolved time interval $[t_b, t_c]$ (with $t_b < t_c$). In doing so, the classifier can use present information to predict future links. The input of this classifier is a feature matrix X, which is based on a network $G_{[t_a,t_b]} = (V_{[t_a,t_b]}, E_{[t_a,t_b]})$ with $E_{[t_a,t_b]} = \{(u, v, t_i) \in E : t_a \leq t_i \leq t_b\}$ and $V_{[t_a,t_b]}$ the nodes taking part in these edges. The feature vector is calculated for each candidate node pair that is not linked (yet) in $G_{[t_a,t_b]}$: $X_{[t_a,t_b]} = \left(V_{[t_a,t_b]} \times V_{[t_a,t_b]}\right) \setminus E_{[t_a,t_b]}$. To ensure that all features are well-defined, we consider only pairs of nodes where both nodes are in the GC of $G_{[t_a,t_b]}$. The **target** of the classifier, y, denotes for a node pair whether a link is present in the evolved network:

$$y_{u,v} = \begin{cases} 0 \text{ if } (u, v, t_i) \notin E \\ 1 \text{ if } (u, v, t_i) \in E \end{cases} \quad \text{for some } t_b < t_i < t_c$$

We note that only the link formation is to be predicted; we do not aim to predict the weight of the link. Accordingly, the prediction can be seen as a *supervised binary classification*.

4.5.2 Features

Below, we explain the composition of the feature vector used for each candidate truck pair (a, b). In Table 4.2, we present *all* 52 features used by the link prediction classifier. The various truck properties will be explained in Subsection 4.5.2A and the spatiotemporal information in Subsection 4.5.2B. All features used can be categorized into four types. We describe each of them in more detail below.

- **Neighborhood features.** These consider relevant operations related to the ego-network (see Section 1.2) properties of the nodes of the candidate pair. The neighborhood of a node is defined by $N(a) = \{v \in V : (a, v, t_i) \in E \text{ for some } t_i\}$. The strength of a node is the summed weight of every link connected to a node, $s_a = \sum_{u \in V} w_{a,u}$.
- **Node features.** These are constructed from information available about the trucks, see Subsection 4.5.2A.
- **Path features.** These relate to the macro-scale properties of the network (Subsection 1.2.4). We consider only the shortest path length in this chapter.
- **Spatiotemporal features.** These relate to the spatial and temporal behavior of the trucks, see Subsection 4.5.2B.

4.5.2A Node features

The ANPR system determines the license plate and country ($country_u$) of each truck u passing by the system. We use $\mathcal{D}_u$ to denote all registrations x_i available of truck u (as explained in Subsection 4.4.1). The registration systems are also equipped with sensors to measure the length ($length_i$), mass ($mass_i$), and the number of vehicle axes ($axes_i$) of each truck. These measurements may slightly differ between registrations. Therefore, we calculate the averages shown in Table 4.3 for each truck in the network.

The $driving\_hours$ and $weekend\_driver$ features are calculated because they are known to vary between trucks operating in different industrial sectors. The actual driving hour t_i (h) is subtracted by $12\,\mathrm{h}$ and the absolute value is taken, such that it is a measure whether a truck u drives at day (resulting in low values for $driving\_hours_u$, or night resulting in high values for $driving\_hours_u$).

4.5.2B Spatiotemporal features

The spatial-temporal features aim to capture the truck pair's spatial and temporal behavior under consideration. We do so by counting the number of registrations in different periods. We consider periods of one week, one month, and one year. These periods are

Table 4.2: The features (of truck pair a and b) of the link prediction classifier and their importance. (The importance of each feature is calculated using the Gini importance, see Subsection 4.5.3.)

Index	Feature	Type	Feature importance
X_1	$truck\_country\,(a) = truck\_country\,(b)$	node	0.005
X_2	$truck\_axes\,(a) + truck\_axes\,(b)$	node	0.006
X_3	$\lvert truck\_axes\,(a) - truck\_axes\,(b)\rvert$	node	0.008
X_4	$truck\_length\,(a) + truck\_length\,(b)$	node	0.017
X_5	$\lvert truck\_length\,(a) - truck\_length\,(b)\rvert$	node	0.040
X_6	$truck\_mass\,(a) + truck\_mass\,(b)$	node	0.016
X_7	$\lvert truck\_mass\,(a) - truck\_mass\,(b)\rvert$	node	0.030
X_8	$driving\_hours\,(a) + driving\_hours\,(b)$	node	0.016
X_9	$\lvert driving\_hours\,(a) - driving\_hours\,(b)\rvert$	node	0.030
X_{10}	$weekend\_driver\,(a) + weekend\_driver\,(b)$	node	0.014
X_{11}	$\lvert weekend\_driver\,(a) - weekend\_driver\,(b)\rvert$	node	0.019
X_{12}–X_{19}	$last\_week_\ell\,(a+b)$ for $\ell = 1, ..., 8$	spatio-temporal	$0 - 0.027$
X_{20}–X_{27}	$last\_month_\ell\,(a+b)$ for $\ell = 1, ..., 8$	spatio-temporal	$0 - 0.057$
X_{28}–X_{45}	$last\_year_\ell\,(a+b)$ for $\ell = 1, ..., 8$	spatio-temporal	$0.010 - 0.060$
X_{46}	$\lvert N\,(a)\rvert + \lvert N\,(b)\rvert$	neighborhood	0.117
X_{47}	$\Big\lvert \lvert N\,(a)\rvert - \lvert N\,(b)\rvert \Big\rvert$	neighborhood	0.013
X_{48}	$\lvert N\,(a) \cup N\,(b)\rvert$	neighborhood	0.093
X_{49}	$\lvert N\,(a) \cap N\,(b)\rvert$	neighborhood	0.021
X_{50}	$s_a + s_b$	neighborhood	0.056
X_{51}	$\lvert s_a - s_b\rvert$	neighborhood	0.017
X_{52}	shortest path length in G	path	0.111

Table 4.3: Overview of available truck information.

Property		Description	Type
$truck\_country_u$		country of registration	string
$truck\_axes_u$	$\underset{x_i \in \mathcal{D}_u}{\text{Median}}\ axes_i$	number of axes	$\mathbb{Z}$
$truck\_length_u$	$\underset{x_i \in \mathcal{D}_u}{\text{Median}}\ length_i$	length	$\mathbb{R}$
$truck\_mass_u$	$\underset{x_i \in \mathcal{D}_u}{\text{Median}}\ mass_i$	mass	$\mathbb{R}$
$driving\_hours_u$	$\underset{x_i \in \mathcal{D}_u}{\text{Mean}}\ \lvert t_i(\text{h}) - 12\text{h}\rvert$	usual driving hours	$[0, 12]$
$weekend\_driver_u$	$\underset{x_i \in \mathcal{D}_u}{\text{Mean}} \begin{cases} 0 \text{ if } t_i = \text{weekday} \\ 1 \text{ if } t_i = \text{weekend} \end{cases}$	fraction driving in weekend	$[0, 1]$

chosen to cover a broad window of possible relevant periods. As an example, for feature $last\_day_\ell\,(a+b)$ registrations are counted for trucks a and b at location ℓ at the last day before the considered time.

4.5.3 Classifier

A *random forest* classifier is used to do link prediction. We choose this classifier because random forests are known to generalize well on unseen data. Our task is to determine the importance of each feature [42, 76].

We now discuss the setup of the classifier. The random forest classifier we used contains 128 decision trees. Larger values usually bring no significant performance gain [140]. Each decision tree is trained on a randomly drawn selection of variables. The number of randomly drawn features equals the square root of the total number of variables, a typical value used in classification [162].

Random sampling with replacement from the data *increases randomness* for each decision tree. The splitting criteria of the nodes are determined by considering the Gini impurity reduction as discussed in [76]. The random forest classifier allows obtaining the *feature importance* by determining the Gini impurity reduction for splitting nodes with a certain feature [76]. We recall that the feature importance is essential, as it enables us to understand the network dynamics by predicting new truck co-driving behavior.

Subsequently, we use the out-of-bag sample of each tree to estimate the performance of the random forest [76, 162]. We then assess the optimal value for the depth of the decision trees in the random forest. The classifier's performance is calculated on the test set, which is a 10% random sample of the data only used for this purpose.

4.5.4 Class imbalance

It is well-known that real-world network link prediction classifiers come with a large class imbalance [196], caused by sparseness of edges (see Subsection 4.4.4). The performance of the random forest classifier may drop if there is a large class imbalance. To overcome this limitation, we use the following two measures.

1. We adjusted the weights of the positive instances so that the total weight of the positive and negative samples are equal.
2. We consider only truck pairs where both trucks are involved in co-driving events in the last two months before time τ. It will reduce the number of considered truck pairs. The class imbalance is also reduced because many truck pairs registered recently have a higher probability of co-driving.

4.6 Experimental setup and results

The setup of the experimental parameters are briefly discussed in Subsection 4.6.1. The results of the link prediction classifier are discussed in Subsection 4.6.2.

4.6.1 Experimental parameter setup

We set the value of τ such that half of the edges are formed. We experimentally found that with this value of τ, the class imbalance is reduced while ensuring that at least a thousand truck pairs are present that will link. The class imbalance is $1 : 61,000$, meaning there is one positive instance for every 61,000 negative instances. Taking the two measures noted in Subsection 4.5.4 reduces the class imbalance to $1 : 15,000$, which improves link prediction performance. Nevertheless, even with this parameter setup, it is still a highly imbalanced set of instances.

Furthermore, we found an optimal maximum depth of three for the decision trees in the random forest using out-of-bag sampling (see Subsection 4.5.3).

For reproducibility purposes, we mention that the random forest is used as implemented in Python sci-kit learn 0.21.2 [149].

4.6.2 Results

We report the trade-off between true and false positives to assess the classifier's accuracy. The relation between these two values is shown in Figure 4.6 using the well-known Receiver Operating Characteristic (ROC) curve [140]. The AUC is 0.84, meaning the classifier can accurately predict whether links will appear. The performance is sufficiently high, and therefore, we continue with the analysis of the feature importance observed.

In Table 4.2, the feature importance is presented. The features are shown for each of the four types (neighborhood, node, path, and spatiotemporal features) in Figure 4.7.

We observe that the neighborhood feature (X_{46}) scores highest with a feature importance of 0.117, closely followed by the single path feature (X_{52}) with an importance of 0.111. The two neighborhood features with the highest scores are X_{46} and X_{48}, with an importance of 0.117 and 0.093, respectively. These features provide the sum of the node pairs' degrees and the union of their neighborhoods, respectively. Both the spatiotemporal and node features score lower, with a maximum feature importance of only 0.060 and 0.040, respectively.

Since the features based on network metrics (neighborhood and path) have higher feature importance, we may conclude from our experiments that the network view (i.e., the structure of the data in the network) on the data is helpful.

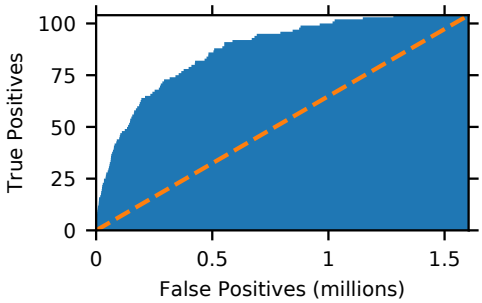


Figure 4.6: The ROC curve of the random forest link prediction classifier.

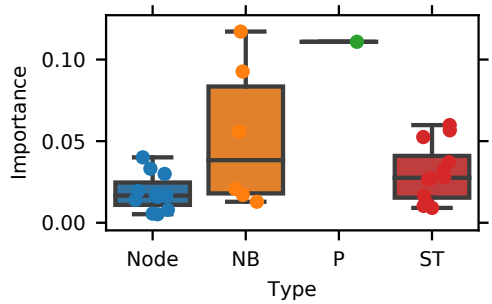


Figure 4.7: The Gini feature importance of the various feature sets. (NB, P, and ST are the neighborhood, path, and spatiotemporal features, respectively.)

4.7 Chapter conclusions and outlook

In this chapter, we addressed Research question 3: "How do network structure and vehicle attributes relate to co-driving behavior?" We compared four sets of features in a link prediction model applied to the co-driving network. By comparing the importance of the different types of features, we observe different abilities in predicting new links. From our experiments, we may conclude that features based on network measures, particularly the neighborhood feature and path feature to a lesser extent, can explicate the dynamics of the studied co-driving network. This means that the network perspective we have adopted in analyzing the spatiotemporal dataset of truck co-driving in the Netherlands has seriously contributed to our comprehension of co-driving behavior. Our second conclusion is that the link prediction approach is a viable method for analyzing spatiotemporal datasets that contain social behavior. Our answer to Research question 3 reads: "The network structure, and especially the ego-network structure of the nodes, relate strongly to co-driving behavior. The same is the case for spatiotemporal information about the truck itineraries. Vehicle attributes show a smaller relation to co-driving behavior."

Chapter outlook

An exciting angle for future work is to use a similar approach to predict which nodes will turn inactive, i.e., will not form any new links. It will result in a substantially smaller set of candidate nodes for the link prediction algorithm. Finally, future work could focus on interpreting and applying the knowledge gained to actually stimulate co-driving behavior, which may in turn facilitate reductions in the fuel use of trucks.